Regresión Lineal Múltiple

Bibliografía:

- Chaterjee, S.; Hadi, A.; Price, B. "Regression Analysis by Example". Wiley
- Montgomery, D.; Peck, E.; Vining, G. "Introducción al Análisis de Regresión Lineal". 3a ed.
- Draper, N.R.; Smith H. (1981) "Applied Regression Analysis". 2nd ed. Wiley N.Y.

Guía de ruta:

- Modelo RLM
- Estimación de mínimos cuadrados
- Supuestos del modelo RLM
- Inferencia en el modelo RLM
- Verificando los supuestos
- Diagnósticos de influencia
- Multicolinealidad
- RLM con variables dummies
- Transformaciones
- Métodos de selección de variables

El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones está dado por:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + \varepsilon_i \quad \text{para i = 1,2,…,n}$$

en forma matricial :

$$\mathbf{Y} = \mathbf{X\beta} + \varepsilon$$

donde

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Observaciones:

- El coeficiente de regresión poblacional β_j, con $j=1,...p$, se llama también "parcial" dado que indica el cambio esperado en la variable de respuesta Y cuando la variable predictora X_j cambia en una unidad adicional **asumiendo que las otras variables predictoras permanecen constantes.**
- El modelo de regresión **"lineal"** requiere linealidad en los parámetros (β's) no necesariamente en los regresores. Entonces un modelo polinomial es un ejemplo de modelo de regresión lineal.

Supuestos del modelo

1. Linealidad
2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
3. $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$
4. Los errores no están correlacionados
5. El vector $\boldsymbol{\varepsilon}$ se distribuye Normal $(\mathbf{0}, \sigma^2 \mathbf{I}_n)$
6. Las variables regresoras no son colineales.

Donde:

- $\boldsymbol{\varepsilon}$ es un vector columna aleatorio de dimensión n.
- $\mathbf{I}_n$ es la matriz identidad de orden n.

Estimación del vector de parámetros β por Cuadrados Mínimos (MCO)

Se tiene que minimizar la suma de cuadrados de los errores.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Haciendo operaciones con los vectores y matrices

$$\mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y} - \mathbf{Y'X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} = \mathbf{Y'Y} - 2\boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta}$$

Derivando con respecto a β e igualando a cero se obtiene el sistema de **ecuaciones normales**

$$\mathbf{X'X}\boldsymbol{\beta} = \mathbf{X'Y}$$

Los estimadores de β son la solución de estas ecuaciones:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Ecuaciones normales (versión no matricial)

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\
&\vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_i \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}
$$

Modelo ajustado

El modelo de regresión ajustado es

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Y los residuales correspondientes pueden también escribirse en un vector

$$\mathbf{e} = \mathbf{Y} - \mathbf{Y}$$

Estimación de la varianza del error

Como en el modelo RLS, la suma de cuadrados de los residuales nos permite estimar la varianza del error, esto es:

$$SSE = \sum_{i=1}^{n} (y_i - y_i)^2 = \sum_i e_i^2 = \mathbf{e}'\mathbf{e}$$

y se tiene un estimador insesgado de la varianza de error:

$$\sigma^2 = \frac{SSE}{n-p-1} = MSE \text{ (cuadrado medio residual)}$$

(ojo! hay p+1 coeficientes)

Ejemplo:

Los datos corresponden a una encuesta de satisfacción de empleados con sus supervisores (Chatterjee). Cada variable Xi suma las respuestas de los empleados de cada uno de los 30 dptos.

Las variables son:

- Y: puntaje de cada supervisor (mide la calidad del trabajo de este)
- X1: resuelve quejas de los empleados
- X2: no permite privilegios
- X3: da oportunidad de aprender
- X4: da aumentos basados en el rendimiento
- X5: es demasiado critico al bajo rendimiento
- X6: tasa de avance a mejor trabajo

Datos (Chaterjee):

Y	X1	X2	X3	X4	X5	X6
43	51	30	39	61	92	45
63	64	51	54	63	73	47
71	70	68	69	76	86	48
61	63	45	47	54	84	35
81	78	56	66	71	83	47
43	55	49	44	54	49	34

Etc…….

Regresión de Y sobre X1 y X2:

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	15,3276	7,1602	0,6360	30,0192	2,1407	0,0415
X1	0,7803	0,1194	0,5354	1,0253	6,5362	<0,0001
X2	-0,0502	0,1299	-0,3167	0,2164	-0,3861	0,7025

el modelo ajustado es

$$y = 15.3276 + 0.7803x_1 - 0.0502x_2$$

Algunos resultados importantes:

1) $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

2) $\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{Y} = \mathbf{H}\mathbf{Y}$

- H se llama “matriz Hat”
- es simétrica e idempotente
- rango(H) = $p+1$ = rango(X)
- H proyecta sobre el espacio columna de X
- Los elementos diagonales de H son los *leverages:*
  - Dicen cuánto pesa y_i en la predicción correspondiente.
  - Dan distancias del punto en X al centroide

Algunos resultados importantes:

3) $\mathbf{e} = \mathbf{Y} - \mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

4) $\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$

o sea los e_i NO son independientes!!

de donde: $\mathrm{Var}(\mathbf{e}_i) = \sigma^2(1\text{-}h_{ii})$

5) $\mathrm{SSE} = \mathbf{e}'\mathbf{e} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$

Coeficiente de determinación

- El coeficiente de determinación se define:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{(Y - 1\bar{y})'(Y - 1\bar{y})}{(Y - 1\bar{y})'(Y - 1\bar{y})} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Indica proporción de variabilidad explicada por el modelo

Problema: aumenta siempre que agregamos regresores.

- El coeficiente de determinación ajustado es más adecuado ya que sólo aumenta si disminuye el cuadrado medio resid. respecto del total.

$$\bar{R}_{aj}^{\,2} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} = 1 - \frac{\sigma^2}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Algunas cuestiones con R^2

- El R^2 aumenta si agrego una v. regresora Z a la ya existente X:

$$R_{XZ}^{\ 2} = R_X^{\ 2} + (1 - R_X^{\ 2})\tilde{r}_{yz}^{\ 2} \geq R_X^{\ 2}$$

$\tilde{r}_{yz}$ = correlación parcial de Y con Z =

= correlación simple entre $e_{y/x}$ y $e_{z/x}$

- Sin embargo $R_{aj}^{\ 2}$ puede disminuir al agregar una variable (puede ser negativo también).
- R^2 es invariante a transformaciones lineales de las variables (en particular, a tipificación o estandarización de los datos).
- R^2 corresponde al cuadrado de la correlación entre Y e Y^

Propiedades de los estimadores MCO

- $\widehat{\boldsymbol{\beta}}$ es insesgado, o sea $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, esto es $E(\beta_j) = \beta_j$

- $\cdot Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1}$, en particular $Var(\beta_j) = \sigma^2 C_{jj}$

Donde C_{jj} es el elemento correspondiente en la inversa de (X`X)

- El estimador minimo cuadrático $\widehat{\boldsymbol{\beta}}$ es el "mejor" estimador dentro de los estimadores lineales insesgados de $\boldsymbol{\beta}$. (Teo. Gauss-Markov)

Inferencia sobre los parámetros

Podemos hacer diferentes pruebas de hipótesis:

- Probar si un coeficiente particular del modelo es = 0.
- Probar si es significativa la regresión, esto es si hay relación lineal entre la v. respuesta y las regresoras.
- Probar si algún grupo de coeficientes es = 0.
- Probar si una combinación lineal de los coeficientes = 0 u otro valor (contraste).

Para hacer inferencia se necesita conocer las distribuciones de los estadísticos asociados:

Bajo el supuesto de Normalidad de los errores, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I_n})$, se tiene que:

$$\frac{SSE}{\sigma^2} \sim \chi^2_{(n-p-1)} \quad \text{y} \quad \boldsymbol{\beta} \sim N(\boldsymbol{\beta}, \sigma^2 (X'X)^{-1})$$

Además β_j y σ^2 independientes,

entonces puedo construír test t y F

Prueba de hipótesis acerca de un coeficiente de regresión individual

- Ho: $\beta_i = 0$
- Ha: $\beta_i \neq 0$

La prueba estadística es la prueba t:

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{\sigma^2 C_{ii}}} \sim t_{n-p-1}$$

Donde, C_{ii} es el i-ésimo elemento de la diagonal de $(X'X)^{-1}$.

Con IC: se rechaza H0 si el IC de este coeficiente no contiene a 0.

Ver regiones de confianza.

Ver ejemplo supervisor.

Tabla de Análisis de Varianza

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados Medios	F
Regresión	SSR	p	MSR=SSR/p	MSR/MSE
Error	SSE	n-p-1	MSE=SSE/n-p-1	
Total	SST	n-1		

Prueba de significancia de la regresión.

H_0: $\beta_1=\beta_2=\ldots=\beta_p=0$

H_1: al menos uno de los coeficientes

es distinto de cero.

Esto corresponde al test F de la tabla de ANOVA y corresponde a ensayar la hipótesis "*la regresión es significativa*".

Además, existe relación entre el estadístico F de esta prueba y el coef. de determinación R^2:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{R^2 / p}{(1-R^2)/(n-p-1)}$$

Un valor grande de F indica rechazar H_0 (la regresión es significativa).

Prueba de hipótesis sobre un grupo de coeficientes

Ho: $\beta_{(q+1)}=\ldots=\beta_p=0$.

H1: Al menos uno de ellos no es cero.

La prueba de F parcial se calcula por:

$$\frac{\frac{SSE(\text{Red})-SSE(\text{Comp})}{p-q}}{\frac{SSE(\text{Comp})}{n-p-1}}=\frac{\frac{SSE(\text{Red})-SSE(\text{Comp})}{p-q}}{MSE(\text{Comp})}\sim F(p-q,n-p-1)$$

SSE(Red) = suma de cuadrados ajustando el modelo **sin** los coef sobre los que se testea (mod. Reducido)

SSE(Comp) = suma de cuadrados ajustando el modelo con todos los coeficientes (mod. Completo)

Ejemplo: datos del supervisor

Variable	N	R^2	R^2 Aj
Y	30	0,7326	0,6628

Coeficientes de regresión

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	10,7871	11,5893	-13,1871	34,7613	0,9308	0,3616
X1	0,61332	0,1610	0,2802	0,9462	3,8090	0,0009
X2	-0,0731	0,1357	-0,3538	0,2077	-0,5382	0,5956
X3	0,3203	0,1685	-0,0283	0,6689	1,9009	0,0699
X4	0,0817	0,2215	-0,3764	0,5399	0,3690	0,7155
X5	0,0384	0,1470	-0,2657	0,3425	0,2611	0,7963
X6	-0,2171	0,1782	-0,5857	0,1516	-1,2180	0,2356

Modelo ajustado:

$$y = 10.787 + 0.613x_1 - 0.073x_2 + 0.32x_3 + 0.08x_4 + 0.038x_5 - 0.217x_6$$

Ej supervisor:

queremos ensayar Ho: $\beta_3=\beta_4=\beta_5=\beta_6=0$

Modelo reducido (solo x1 y x2)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,826[a]	,683	,660	7,10207

a. Variables predictoras: (Constante), X2, X1

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2935,103	2	1467,551	29,095	,000[a]
	Residual	1361,864	27	50,439		
	Total	4296,967	29			

a. Variables predictoras: (Constante), X2, X1

b. Variable dependiente: Y

Ej supervisor

Modelo completo (todas las predictoras)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,856[a]	,733	,663	7,06799

a. Variables predictoras: (Constante), X6, X1, X5, X2, X3, X4

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3147,966	6	524,661	10,502	,000[a]
	Residual	1149,000	23	49,957		
	Total	4296,967	29			

a. Variables predictoras: (Constante), X6, X1, X5, X2, X3, X4

b. Variable dependiente: Y

Ej supervisor

El estadístico F para esta prueba será:

$$F = \frac{\frac{SSE(\text{Red}) - SSE(\text{Comp})}{p-q}}{MSE(\text{Comp})} = \frac{\frac{1361.86 - 1149.00}{4}}{49,957} = 1,06$$

Que corresponde a un *p-valor* = 0,6 o sea no se puede rechazar H_0

(F tiene 4g.l.numerador y 23 g.l.denominador)

Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Se desea predecir el valor medio de la variable de respuesta Y para una serie de valores de las variables predictoras $X_1,\ldots\ldots X_p$.

Consideremos el vector de valores observados

$$\mathbf{x}'_o = (1, x_{01}, x_{021}, .., x_{0p})$$

El **valor medio** de la variable de respuesta Y será

$$E(Y_0) = \mathbf{x}'_\mathbf{o}\boldsymbol{\beta} \text{ que se estima con } \hat{y}_o = \mathbf{x}'_\mathbf{o}\hat{\boldsymbol{\beta}}$$

con varianza

$$Var(\hat{y}_o) = \mathbf{x}'_\mathbf{o}\,\mathbf{Var}((\hat{\boldsymbol{\beta}})\mathbf{x}_\mathbf{o} = \sigma^2\mathbf{x}'_\mathbf{o}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\mathbf{o}$$

Se asume que los errores están normalmente distribuidos.

Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Un intervalo de confianza para el valor medio de Y dado que **x=x'$_o$** es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)} s\sqrt{\mathbf{x'}_o(\mathbf{X'X})^{-1}\mathbf{x}_o}$$

Un intervalo de predicción para el valor individual de Y dado **x=x$_o$** es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)} s\sqrt{1+\mathbf{x'}_o(\mathbf{X'X})^{-1}\mathbf{x}_o}$$

Leverage del "punto" x0

Verificando los supuestos

Recordemos que el modelo RLM supone:

1. La relación y versus regresoras es (aprox) lineal
2. La media de los errores ε es 0
3. La Var(ε) es constante e igual a σ^2
4. Los errores no están correlacionados
5. Los errores ε se distribuyen de modo Normal
6. los regresoras no son colineales

Queremos detectar violaciones de estos supuestos a partir del análisis de los residuales del modelo. Estos muestran:

- desviación entre datos y ajuste
- variabilidad no explicada por el modelo
- valores observados de los errores

Verificando los supuestos

A partir de las propiedades se ve que, si el modelo es correcto, los residuales no se correlacionan con los predichos. Por esto se puede verificar los supuestos graficándolos en el plano.

Indicadores gráficos

- Antes de ajustar modelo: histogramas de la variable respuesta, gráficos de dispersión con correlaciones

  (Y ~x_i), etc

- Después: gráficos de residuales

  - Para chequear linealidad y normalidad, homogeneidad de varianzas
  - Para detectar outliers

  Para que sean visibles más claramente, se puede ver los residuales 'escalados' de varias maneras.

Residuales Estandarizados

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad \text{y} \quad \mathbf{Var(e)} = \sigma^2(\mathbf{I} - \mathbf{H})$$

$$E(e_i) = 0 \qquad V(e_i) = \sigma^2(1 - h_{ii})$$

- Entonces una forma de escalarlos es

$$d_i = \frac{e_i}{\sqrt{\sigma^2}} = \frac{e_i}{\sqrt{MSE}}$$

Tienen media 0 y varianza aproximada =1.

Por lo tanto valores d*i* por encima de 3 ó debajo de -3 son indicativos de datos atípicos.

Residuales Estudentizados

Se definen por

$$r_i = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

Residuales Estudentizados **externamente**

Se definen por

$$r_i^* = \frac{e_i}{\sigma_{(i)}\sqrt{1-h_{ii}}}$$

$\sigma^2_{(i)} =$ obtenidos excluyendo la observ. i

Residuales PRESS

Se definen por

$$e_{(i)} = y_i - y_{(i)} = \frac{e_i}{1 - h_{ii}}$$

Tienen media 0 y varianza aproximada $\frac{\sigma^2}{1 - h_{ii}}$

Sirven para medir la capacidad de predicción del modelo porque identifican observaciones que afectan a la predicción por sobre los demás.

Si h_{ii} es grande y $e_{(i)}$ también, este residual PRESS difiere mucho del residual simple lo que indica un punto de gran influencia en la predicción.

Outliers y puntos influyentes

Una observación es considerado un **outlier** si está alejado de la mayoría de los datos sea en la dirección de Y o en la de alguno de los regresores ó en ambos.

Diferenciamos outliers en Y (se los ve en los gráficos de residuales) y outliers en X's.

Una observación es considerado un **valor influyente** si su presencia afecta de manera importante el comportamiento del modelo. Por ejemplo en el caso de regresión simple remover un valor influencial cambiaría mucho el valor de la pendiente.

Medidas para detectar datos influyentes

Leverages: h_{ii} (elementos diagonales de H)

- Para cada observación mide la distancia de la observación al centro de la media de todas las observaciones de las variables independientes. (Leverage alto = punto alejado en el espacio de las X)
- Valores altos en la diagonal pueden dar mucho peso a la predicción del valor de la variable dependiente (ver PRESS).
- El rango de valores es de 0 a 1, con media *p/n*, *p* es el número de predictores y *n* es el tamaño de muestra. Si supera *2p/n* puede considerarse atípico.

Medidas de influencia: Distancia Cook

Mide el cambio que ocurriría en el vector de coeficientes estimados de regresión si la i-ésima observación fuera omitida.

Se calcula como:

$$CD_i^2 = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{ps^2} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} = r_i^2 \frac{h_{ii}}{p(1-h_{ii})}$$

donde $\hat{\mathbf{y}}_{(i)}$ es el vector predicho si no se usa la observación i-ésima

$r_i =$ resid estudentiz.

Hay distintos criterios para el valor de corte, algunas sugieren que una observación sea considerada como **influencial** si CD_i^2 es mayor a *4/n*, ó a *1*. En general, se recomienda analizar las de mayor valor que queden alejadas del resto. (Draper, pag 212)

Medidas de influencia: DFBetas

DFBeta(j,i) mide cuánto cambiaría el coef estimado de βj si se omitiera la observación i-ésima.

$$(DFBETAS)_{ji} = \frac{\beta_j - \beta_{j,(i)}}{s_{(i)}\sqrt{c_{jj}}}$$

Donde *cjj* es el *j*-ésimo elemento de la diagonal de la inversa de **(X'X)**

Si |DFBETAS|ji > 2/raiz(n) para algun j entonces la i-esima observacion es posiblemente un valor influencial.

Medidas de influencia: DFFITS

$$DFFITS_i = \frac{(\hat{y}_i - \hat{y}_{(i)})}{\sqrt{h_{ii}.s^2_{(i)}}}$$

Un criterio es $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ indica un posible valor influencial.

Notar que:

$$CD_i = \frac{\sigma^2_{(i)}}{p\sigma^2} DFFITS_i^2$$

DFFIT mide cuánto cambiaría el valor predicho si se omitiera la observación i-ésima.

Obs	y	x1	x2	COOK	Leverage	COVRATIO	DFFit	DFB0	DFB1	DFB2
1	16,68	7	560	0,10009	0,0618	0,87108	-0,56988	-0,19715	0,06741	-0,00151
2	11,5	3	220	0,00338	0,0307	1,21492	0,08722	0,10049	-0,00832	0,00005
3	12,03	3	340	0,00001	0,05873	1,27568	-0,00545	-0,00395	0,00069	-0,00001
4	14,88	4	80	0,07765	0,04537	0,876	0,45966	0,47771	0,01453	-0,00095
5	13,75	6	150	0,00054	0,03501	1,2396	-0,03604	-0,03554	-0,00232	0,00009
6	18,11	7	330	0,00012	0,00287	1,19991	-0,01297	-0,01648	0,00031	0
7	8	2	110	0,00217	0,0418	1,23975	0,07524	0,08749	-0,00389	-0,00004
8	17,83	7	210	0,00305	0,02373	1,20564	0,07872	0,07968	0,00582	-0,0002
9	79,24	30	1460	3,41932	0,45829	0,34221	7,36919	-2,10601	0,11822	0,00406
10	21,5	5	605	0,05385	0,1563	1,3054	0,58041	0,11931	-0,0582	0,00124
11	40,33	16	688	0,0162	0,04613	1,17173	0,21089	-0,03802	0,01598	-0,00001
12	21	10	215	0,0016	0,07366	1,29061	-0,07605	-0,03395	-0,0085	0,0002
13	13,5	4	255	0,00229	0,02112	1,20705	0,06686	0,08104	-0,00621	0,00004
14	19,75	6	462	0,00329	0,03824	1,22768	0,09062	0,05544	-0,01169	0,00023
15	24	9	448	0,00063	0,00111	1,19185	0,02878	0,02498	-0,00084	0,00003
16	29	10	776	0,00329	0,12594	1,36922	-0,13189	-0,00302	0,01125	-0,00031
17	15,35	6	200	0,0004	0,01943	1,21925	0,02757	0,03238	0,00113	-0,00006
18	19	7	132	0,04398	0,05626	1,06921	0,36732	0,27105	0,03221	-0,00098
19	9,5	3	36	0,01192	0,05645	1,21525	0,19141	0,19222	0,00409	-0,00036
20	35,1	17	770	0,13244	0,06168	0,75982	-0,65517	0,17293	-0,03444	-0,00032
21	17,9	10	140	0,05086	0,12528	1,23769	-0,51761	-0,17856	-0,05101	0,00122
22	52,32	26	810	0,45105	0,35158	1,39808	-2,37261	0,42549	-0,17042	0,00202
23	18,75	9	450	0,0299	0,00126	0,88968	-0,19829	-0,17073	0,0062	-0,00019
24	19,83	8	635	0,10232	0,08061	0,94763	-0,62717	-0,12737	0,06702	-0,00163
25	10,75	4	150	0,00011	0,02664	1,2311	-0,01518	-0,01887	0,00015	0,00002

Ejemplo Tiempos de Entrega (Montgomery)

modelo	Beta_0	Beta_1	Beta_2	MSE	R^2
Con todos	2.341	1.616	0.014	10.624	0.9596
Sin caso 9	4.447	1.498	0.010	5.905	0.9487
Sin caso 22	1.916	1.786	0.012	10.066	0.9564
Sin 9 y 22	4.643	1.456	0.011	6.163	0.9072

Autocorrelación en los errores

- Cuando la variable predictora es tiempo, pudiera ocurrir que los errores estén autocorrelacionados
- La prueba de Durbin Watson permite detectar si hay una positiva correlación serial de orden uno.

H_o: $\rho = 0$ vs H_a: $\rho > 0$ $(\rho < 0)$

Valores del estadístico lejos de 2 son indicativos de autocorrelación.

Multicolinealidad

Un conjunto de predictoras X_1, X_2,….X_p son colineales si existen constantes c_o,c_1,…..c_p, tales que vale

$$\sum_j c_j X_j = c_o$$

Cuando se da esta relación exacta, diremos que el modelo está mal especificado.

Cuando hay dependencias (casi) lineales entre los regresores se dice que existe un problema de **multicolinealidad**.

Fuentes de multicolinealidad

- El método de recolección de datos.
- Restricciones en el modelo o en la población.
- Sobredefinición del modelo

Detectando multicolinealidad

Consideremos el modelo escalado y centrado

$$Y^* = \beta^*_1 X^*_1 + \beta^*_2 X^*_2 + ... \beta^*_p X^*_p + \varepsilon$$

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}} \quad ; \quad y_i^* = \frac{y_i - \bar{y}}{\sqrt{s_{yy}}}$$

Si X* tiene columnas dependientes, entonces X también las tiene. La recíproca no vale.

Ecuaciones normales

Quedan

$$(X`*X*)\boldsymbol{\beta}* = X`*Y*$$

donde

$$X`*X* = \begin{bmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & r_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1} & r_{p2} & . & . & 1 \end{bmatrix}$$

r_{ij} representa la correlación entre X_i y X_j

Efectos de multicolinealidad

- Los estimadores de los coeficientes tienen grandes desvíos, lo cual afecta inferencia y predicción.
- Es poco acertado dar la interpretación usual a los coeficientes de la regresión ajustada (por ejemplo dan con signo contrario al esperado).
- Los valores estimados para los coeficientes son sensibles a pequeños cambios en los datos o a quitar/agregar alguna variable.
- Los coeficientes estimados pueden dar muy grandes en valor absoluto.

Factor de inflación de la varianza

$$VIF_j = C^*_{jj} = \frac{1}{1\text{-}R_j^2}$$

Los VIF son los elementos que están en la diagonal de la matriz C.

R_j^2 es el coef de determinación en la regresión de Xj versus las restantes regresoras. Si es cercano a 1 entonces el VIF (o FIV) de ese coeficiente es grande y consecuentemente es grande la varianza.

El VIF representa el incremento en la varianza debido a la presencia de multicolinealidad.

Equivalentemente, se puede mirar el *índice de tolerancia* = 1/VIF*i*

Indice de condición

El *índice de condición* de la matriz correlación X*'X*, la cual es de la forma

$$\begin{bmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & r_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1} & r_{p2} & . & . & 1 \end{bmatrix}$$

Se define como

$$K = \frac{\lambda_{\max}}{\lambda_{\min}}$$

λ son los autovalores de X`*X*

Cómo detectar la multicolinealidad

- Examinar matriz de correlaciones X`X
- VIF
- ver autovalores de X`X

- Calcular el índice condición para cada variable:

$$IC_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}}$$

Y ver que no den mayores de 30.

Cómo detectar la multicolinealidad

- Si hay correlaciones altas en las predictoras (mirarlas en X`X con X la matriz reescalada).
- Si alguno de los factores de inflación de la varianza

  $FIV_i > 5$ (o >10)
- Si el número condición

  $K > 100$: multicolinealidad moderada;

  $K > 1000$: multicolinealidad grave
- Si el test F de la regresión da significativo pero las pruebas t individuales no, puede haber colinealidad.
- Ojo! La multicolinealidad no se detecta con los residuos. No es un error del modelo sino una condición de los datos.

Métodos para remediar el problema de multicolinealidad

Algunas propuestas:

- Recolección de datos adicionales
- Reespecificar el modelo (por ej eliminar variables)
- Regresión Ridge
- Regresión en componentes principales
- PLSR

El problema de multicolinealidad también está relacionado con los métodos de selección de variables y estos pueden ser otra manera de resolver el problema de multicolinealidad.

OJO! los métodos de selección de variables no se recomiendan con datos colineales. En todo caso el que se podría usar es backward…. (pag 291 de Chaterjee)

Datos supervisor: colinealidad?

$$\mathrm{VIF}_1 = 2.7, \quad \mathrm{VIF}_2 = 1.6, \quad \mathrm{VIF}_3 = 2.3,$$

$$\mathrm{VIF}_4 = 3.1, \quad \mathrm{VIF}_5 = 1.2, \quad \mathrm{VIF}_6 = 2.0.$$

$$\lambda_1 = 3.169, \quad \lambda_2 = 1.006, \quad \lambda_3 = 0.763,$$

$$\lambda_4 = 0.553, \quad \lambda_5 = 0.317, \quad \lambda_6 = 0.192.$$

Nº condición = 3.169 / 0.192 = 32…(aprox)